

Aplicações Informáticas na Biomedicina

8.^a Aula Prática Laboratorial

Mestrado Integrado em Engenharia Informática

Ano Letivo 2019/2020

Marisa Esteves

13 de Novembro de 2019



Universidade do Minho

Plano de Aula

1. Contextualização e demonstração do Talend;
2. Resolução da 5.^a ficha prática laboratorial pelos alunos em grupo.

Processo ETL

Definição

O processo ETL (*Extract, Transform, Load*) é um conjunto de processos que inclui a extração de dados de fontes de informação internas e externas, podendo estar em diferentes formatos, a transformação dos dados de acordo com as necessidades da organização e, finalmente, o carregamento dos mesmos numa estrutura de dados, como por exemplo um data mart ou um data warehouse.

Processo ETL

Definição

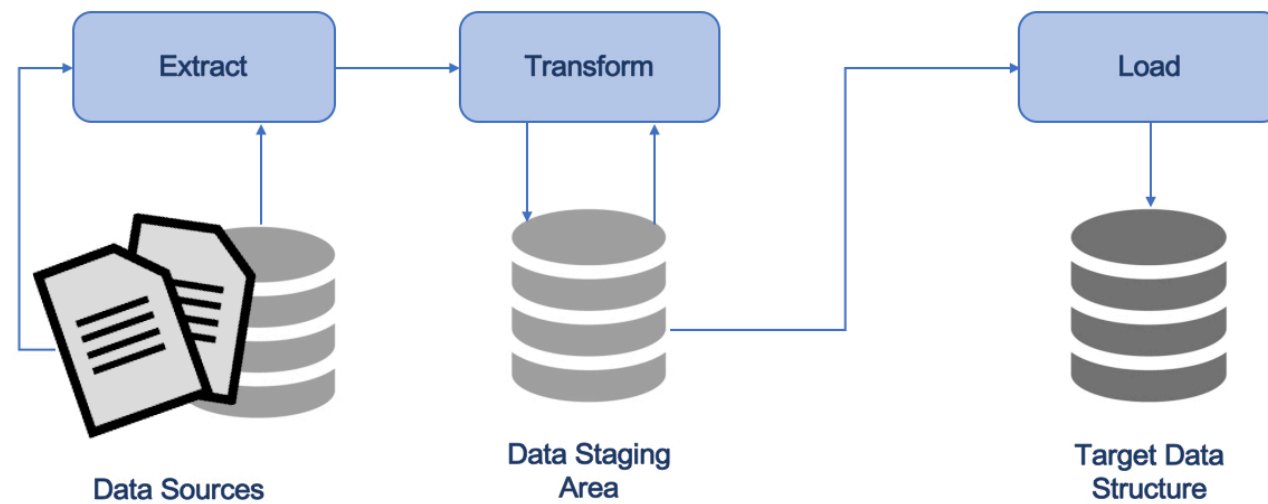


Figura 1 – Esquema do processo ETL.

Processo ETL

Porquê?

Os dados estão espalhados
por diferentes localizações

Os dados estão
armazenados em diferentes
tipos de formato

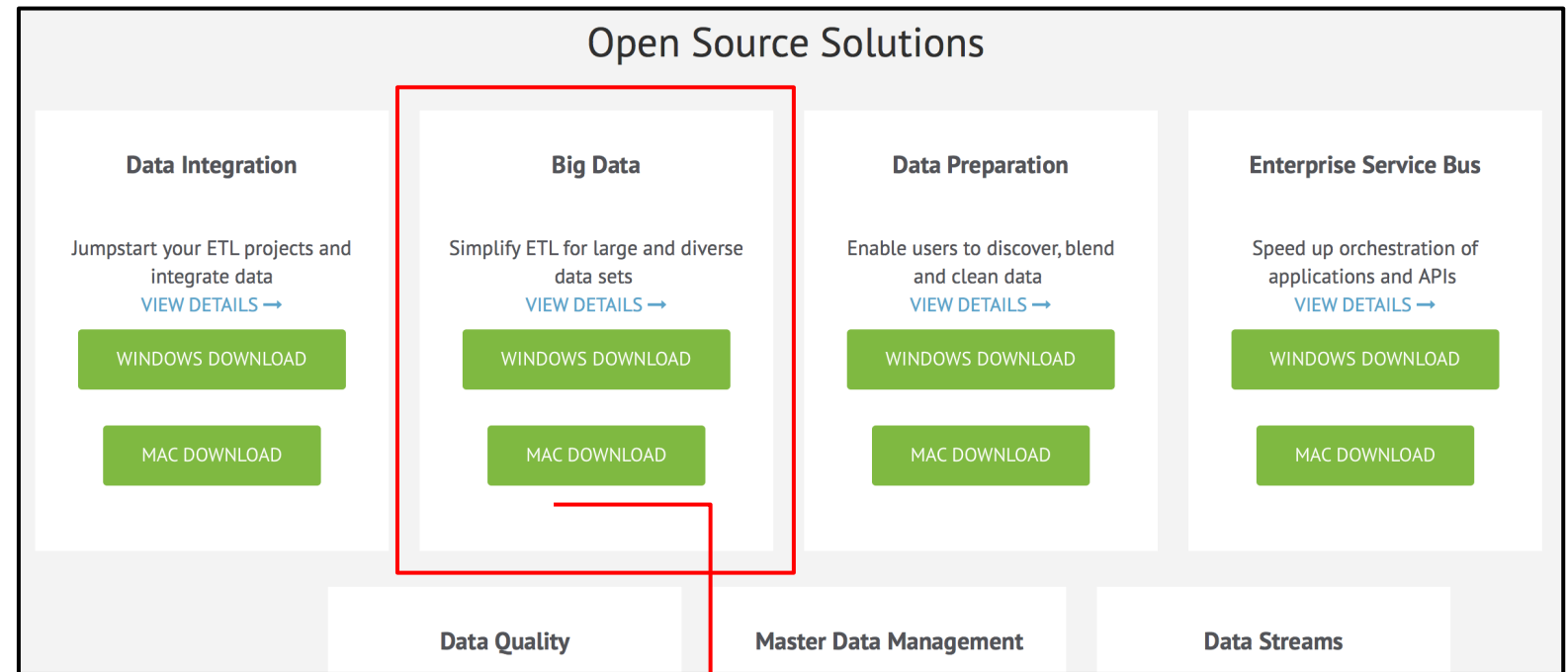
O volume de dados continua
a aumentar

Os dados podem estar
estruturados, semi-
estruturados ou não
estruturados

Instalação

Talend

<https://www.talend.com/products/talend-open-studio/>



+ third party libraries (jars)

Talend

Contextualização

Ferramentas ETL – São ferramentas que combinam as três fases do processo de ETL (*Extract, Transform, Load*) numa única ferramenta.

Fácil de usar

User friendly GUI

Tratamento dos
erros incorporado

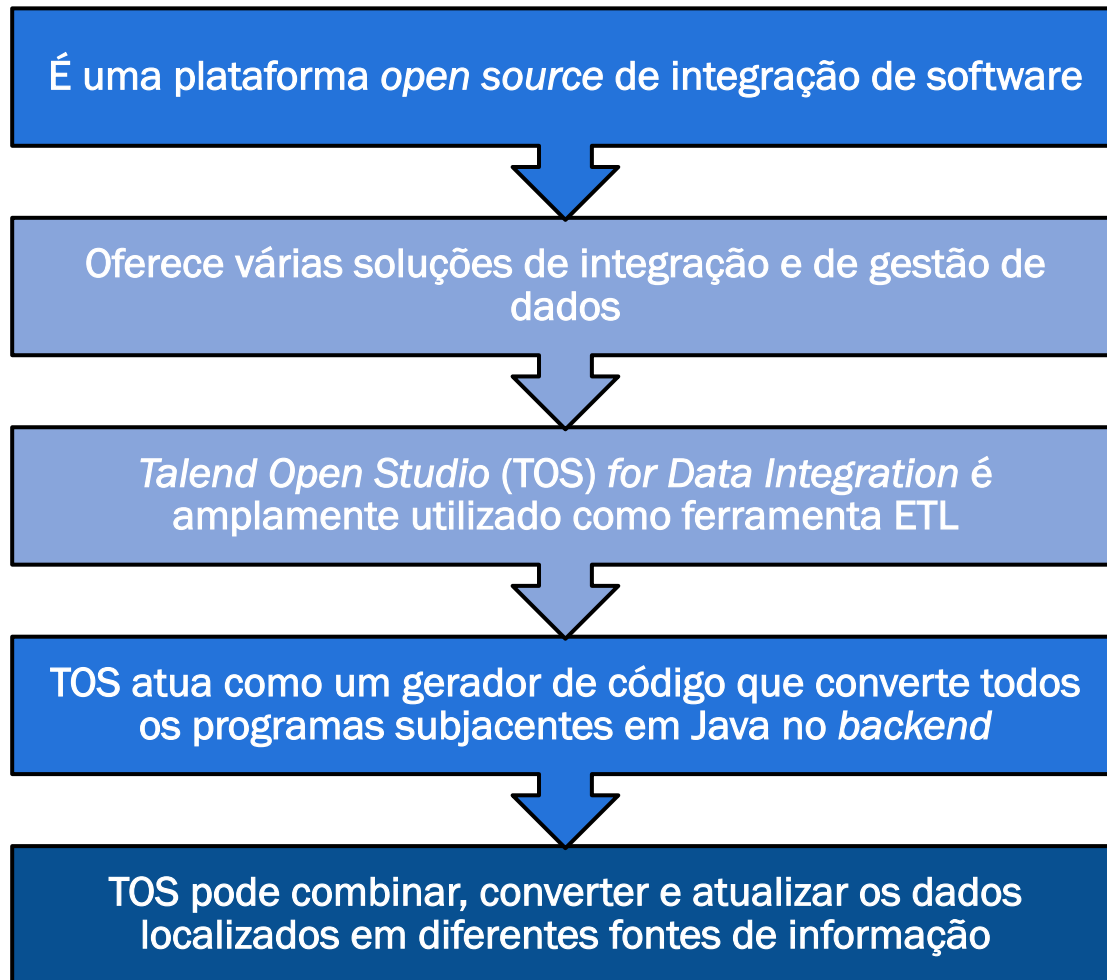
Reduz as despesas

Melhora a gestão
dos dados

Melhora o
desempenho

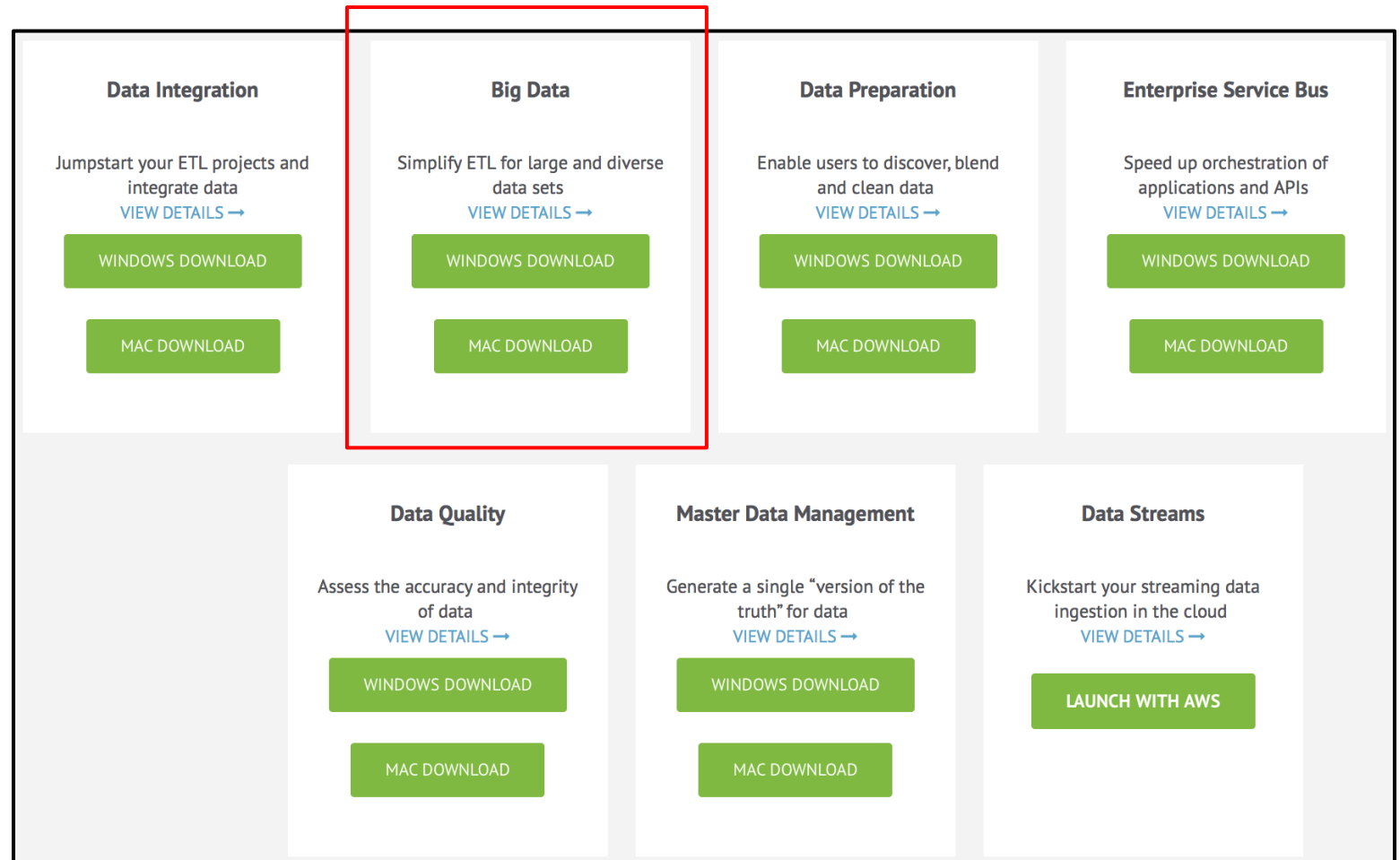
Talend

Definição



Talend

Definição

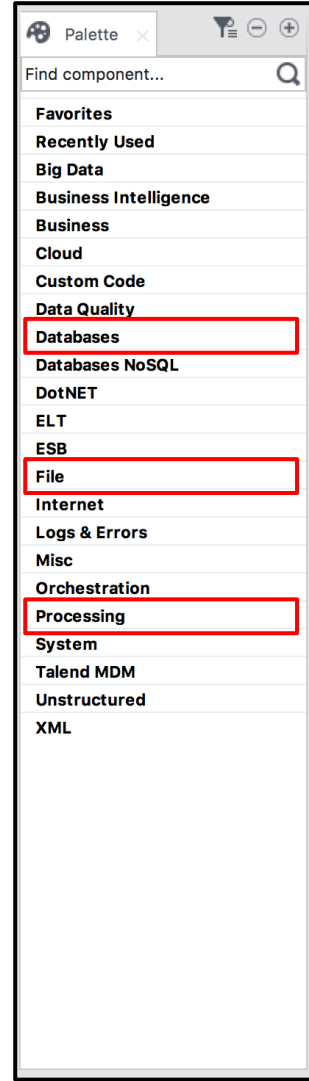


The image shows a grid of seven Talend product categories. The 'Big Data' category is highlighted with a red border. Each category includes a brief description, a 'VIEW DETAILS' link, and download buttons for Windows and Mac. The 'Data Streams' category includes a 'LAUNCH WITH AWS' button instead of download buttons.

Category	Description	VIEW DETAILS →	Windows Download	Mac Download
Data Integration	Jumpstart your ETL projects and integrate data	VIEW DETAILS →	WINDOWS DOWNLOAD	MAC DOWNLOAD
Big Data	Simplify ETL for large and diverse data sets	VIEW DETAILS →	WINDOWS DOWNLOAD	MAC DOWNLOAD
Data Preparation	Enable users to discover, blend and clean data	VIEW DETAILS →	WINDOWS DOWNLOAD	MAC DOWNLOAD
Enterprise Service Bus	Speed up orchestration of applications and APIs	VIEW DETAILS →	WINDOWS DOWNLOAD	MAC DOWNLOAD
Data Quality	Assess the accuracy and integrity of data	VIEW DETAILS →	WINDOWS DOWNLOAD	MAC DOWNLOAD
Master Data Management	Generate a single "version of the truth" for data	VIEW DETAILS →	WINDOWS DOWNLOAD	MAC DOWNLOAD
Data Streams	Kickstart your streaming data ingestion in the cloud	VIEW DETAILS →	LAUNCH WITH AWS	

Talend

Componentes



A ferramenta inclui mais de 900 componentes e conectores divididos em famílias de componentes. As mais utilizadas são as seguintes:

File Components	Processing Components	Databases Components
<ul style="list-style-type: none">• tFileInputDelimited;• tFileInputExcel;• tFileOutputDelimited;• tFileOutputExcel;• tFileList;• tFileExists;• tFileCopy.	<ul style="list-style-type: none">• tMap;• tJoin;• tFilterRow;• tSortRow.	<ul style="list-style-type: none">• tDBConnection;• tDBInput;• tDBRow;• tDBCommit;• tDBOutput.

Demonstração

Talend

The screenshot displays the Talend Studio interface with a job design for 'demo_job1 1.1'. The job is enclosed in a red rectangular frame. The design includes the following components and connections:

- mysql_localhost_talend** (Database Connection) connects to **OnSubjobOk** (Job Trigger).
- OnSubjobOk** connects to **lista_espera_blo** (File Input Delimited).
- lista_espera_blo** connects to **tMap_1** (Map Component).
- tMap_1** has two main outputs:
 - output (Main)** connects to **tDBOutput_3** (Database Output).
 - row2 (Main)** connects to **tFilterRow_1** (Filter Row).
- tFilterRow_1** has two outputs:
 - row3 (Filter order:1)** connects to **tSortRow_1** (Sort Row).
 - row4 (Main)** connects to **tFileOutputExcel_1** (File Output Excel).
- tSortRow_1** connects to **tFileOutputExcel_2** (File Output Excel).
- tFileOutputExcel_1** and **tFileOutputExcel_2** are both labeled as **row4 (Main)**.

The interface also shows a left-hand repository pane with a tree view of the project structure, including 'Business Models', 'Job Designs', 'Contexts', 'Code', 'SQL Templates', and 'Metadata'. The bottom pane shows the 'Job demo_job1' configuration with tabs for 'Basic Run', 'Debug Run', 'Advanced settings', 'Target Exec', and 'Memory Run'. The 'Execution' section includes buttons for 'Run', 'Kill', and 'Clear'.

Resolução da 5.ª Ficha Prática Laboratorial

1 Definição, Criação e Execução de Jobs no Talend

O ficheiro disponibilizado juntamente com esta ficha prática laboratorial, nomeadamente “mental_health.csv”, contém dados reais de um estudo realizado em 2014 que avalia as atitudes dos trabalhadores em empresas de Tecnologias da Informação (TIs) relativamente à saúde mental, bem como a frequência de transtornos mentais em ambientes de trabalho.

O *dataset* é constituído por 27 diferentes colunas, nomeadamente:

- *timestamp*;
- *age*;
- *gender*;
- *country*;
- *state* (*If you live in the United States, which state or territory do you live in?*);
- *self_employed* (*Are you self-employed?*);
- *family_history* (*Do you have a family history of mental illness?*);
- *treatment* (*Have you sought treatment for a mental health condition?*);
- *work_interfere* (*If you have a mental health condition, do you feel that it interferes with your work?*);
- *no_employees* (*How many employees does your company or organization have?*);
- *remote_work* (*Do you work remotely (outside of an office) at least 50% of the time?*);
- *tech_company* (*Is your employer primarily a tech company/organization?*);

Resolução da 5.ª Ficha Prática Laboratorial

- *benefits* (Does your employer provide mental health benefits?);
- *care_options* (Do you know the options for mental health care your employer provides?);
- *wellness_program* (Has your employer ever discussed mental health as part of an employee wellness program?);
- *seek_help* (Does your employer provide resources to learn more about mental health issues and how to seek help?);
- *anonymity* (Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources?);
- *leave* (How easy is it for you to take medical leave for a mental health condition?);
- *mental_health_consequence* (Do you think that discussing a mental health issue with your employer would have negative consequences?);
- *phys_health_consequence* (Do you think that discussing a physical health issue with your employer would have negative consequences?);
- *coworkers* (Would you be willing to discuss a mental health issue with your coworkers?);
- *supervisor* (Would you be willing to discuss a mental health issue with your direct supervisor(s)?);
- *mental_health_interview* (Would you bring up a mental health issue with a potential employer in an interview?);

Resolução da 5.ª Ficha Prática Laboratorial

- *phys_health_interview* (Would you bring up a physical health issue with a potential employer in an interview?);
- *mental_vs_physical* (Do you feel that your employer takes mental health as seriously as physical health?);
- *obs_consequence* (Have you heard of or observed negative consequences for coworkers with mental health conditions in your workplace?);
- *comments* (Any additional notes or comments).

Tendo em conta o seu enquadramento teórico e prático sobre o processo de ETL e o Talend, o objetivo principal a ser alcançado com a resolução desta ficha prática laboratorial é a definição, a criação e a execução de diversas *jobs* com as componentes disponibilizadas na mesma.

Com esta ficha prática laboratorial, pretende-se que:

1. Defina três diferentes *jobs* e justique a utilidade de cada *job* com uma contextualização.
2. Crie as *jobs* definidas no Talend, recorrendo às componentes adequadas. No entanto, deverá incluir, no mínimo, os seguintes elementos:
 - (a) *File Components*: *tFileInputDelimited*, *tFileOutputExcel*;
 - (b) *Processing Components*: *tMap*, *tFilterRow*, *tSortRow*;
 - (c) *Databases Components*: *tDBConnection*, *tDBOutput*;
 - (d) Filtros de rejeição.
3. Corra cada *job* definida e criada.

Resolução da 5.^a Ficha Prática Laboratorial

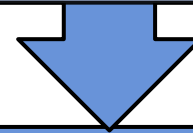
A resolução desta ficha prática laboratorial deverá ser enviada por e-mail para os docentes da unidade curricular - marisa@di.uminho.pt e jmac@di.uminho.pt - até ao dia 19 de novembro de 2019. Um responsável por grupo deverá enviar num mail com título “Ficha5_AIB_GRUPOX”, substituindo “X” pelo número do grupo de trabalho, a pasta zipada do projeto com as *jobs* criadas, bem como um documento de apoio no formato .pdf com a contextualização e a explicação de cada *job* criada (com figuras).

Instalação

Power BI

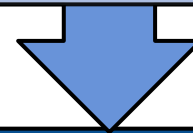
1. Criar uma conta no Power BI com o seu e-mail institucional

<https://powerbi.microsoft.com/en-us/get-started/>



2. Aceder ao Power BI online com a conta criada

<https://app.powerbi.com>



3. *Download* e instalação do Microsoft Power BI Desktop

<https://powerbi.microsoft.com/en-us/downloads/>



+ Ativar o Power BI Pro